Set Mapping Induced Image Perceptual Similarity Distance

En-hui Yang, Xiang Yu, and Jin Meng

Abstract—Perceptual similarity between any two independent images is addressed with a set mapping perspective by first expanding each image \( A \) into a set \( \phi(A) \) of images, and then defining the similarity between any two images \( A \) and \( B \) as the smallest average distortion \( d \) per pixel between any pair of images, one from \( \phi(A) \) and the other from \( \phi(B) \). The resulting similarity metric is dubbed the set mapping induced similarity distance (SMID) between \( A \) and \( B \) and denoted by \( d_s(A,B) \). Several examples of the set mapping \( \phi \) are illustrated; each of them gives a set \( \phi(A) \) of images, which may contain images perceptually similar to \( A \) to certain degree. It is shown that under some mild conditions, \( d_s(A,B) \) is indeed a pseudo distance over the set of all images. Analytic formulas for and lower bounds to \( d_s(A,B) \) are also developed for some set mappings \( \phi \). When compared with other similarity metrics such as those based on histogram-based methods, SIFT, autocorrelagram, etc., the SMID shows better discriminating power on image similarity. Experimental results also show that the SMID is well aligned with human perception for image similarity.

I. INTRODUCTION

Image perceptual similarity is a fundamental issue in image processing, retrieval, management, and understanding in general [1], [2], [3]. As shown in Figure 1, in image processing such as image compression, denoising, and reconstruction, one needs a similarity metric or function to measure the perceptual similarity between the original image \( A \) and a processed image \( \hat{A} \) and evaluate the performance of the corresponding image processing system. In image retrieval and management, one needs a proper similarity metric or function to measure the perceptual similarity between any two independent images \( A \) and \( B \). Once such a similarity metric or function is defined for any two independent images, it can be used to retrieve images in a database which are perceptually similar to a query image according to the similarity metric or function in image retrieval, and to organize images into different groups according to their mutual similarity in image management.

![Image processing diagram](image.png)

Fig. 1. An image processing diagram.

With reference to Figure 1, in image processing, one often defines the similarity between the original image \( A = (a(i,j)) \) and processed image \( \hat{A} = (\hat{a}(i,j)) \), where \( a(i,j) \) and \( \hat{a}(i,j) \), \( 1 \leq i \leq m \), \( 1 \leq j \leq n \), are both real numbers, as the average distortion per pixel:

\[
d(A, \hat{A}) \triangleq \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} d(a(i,j), \hat{a}(i,j)) \quad (1.1)
\]

where \( d(a(i,j), \hat{a}(i,j)) \in [0, +\infty) \) represents the distortion between pixel values \( a(i,j) \) and \( \hat{a}(i,j) \). For example, the similarity function as defined in (1.1) along with the squared error distortion \( d(a(i,j), \hat{a}(i,j)) = (a(i,j) - \hat{a}(i,j))^2 \) is predominantly used in image and video compression [12], [13]. Except for the case where the quality of \( \hat{A} \) is really poor, the similarity function as defined in (1.1) along with the squared error distortion \( d(a(i,j), \hat{a}(i,j)) = (a(i,j) - \hat{a}(i,j))^2 \) is more or less consistent with human perception in the context of image processing. This, together with its simplicity in its computation and rate distortion optimization [26], [27], [25], has contributed to the great success of image and video compression, as evidenced by the evolution of the MPEG-series and H-series video compression standards from MPEG 1/H.261 to HEVC (the newest video coding standard).

In contrast, there is no universally accepted similarity function for measuring the perceptual similarity between any two independent images in image retrieval and management yet [3], [24], [2]. As shown in Figure 2, a typical approach in image retrieval is to first extract features from an image, then derive a signature of the image from the extracted features, and finally determine the similarity between a pair of images based on their respective signatures. Most, if not all, content based image retrieval (CBIR) systems follow more or less these steps; their differences lie in what types of features would be extracted, how a signature would be derived from extracted features, and how distances (i.e., dissimilarity measures) between signatures are defined and calculated [3]. Depending on the mathematical expression of signatures, the Euclidean distance, Hausdorff distance, weighted sum of vector distances, and Kullback-Leibler (KL) divergence (or relative entropy) between signatures have been used [3]. However, no matter how the three steps shown in Figure 2 are executed, one thing is certain: as one moves along the three step chain shown in Figure 2, not only would each step introduce expensive computation, but more importantly the notion of distance between signatures also becomes less intuitive and is increasingly disconnected from the original pair of images.

In this paper, we shall take a different approach to addressing the similarity between any pair of images \( A \) and...
B. Contrary to Figure 2 where images A and B are first “compressed” into their signatures and the similarity between A and B is then defined through the notion of distance between their respective signatures, we instead first expand each image A (B, resp.) into a set \( \phi(A) \) (\( \phi(B) \), resp.) of images, which may contain images perceptually similar to A (B, resp.), and then define the similarity between A and B as the smallest average distortion per pixel between any pair of images, one from \( \phi(A) \) and the other from \( \phi(B) \), through the simple distortion formula defined in (1.1). The resulting similarity value (denoted by \( d_\phi(A, B) \)) will be referred to as the set mapping induced similarity distance (SMID) between A and B. No feature extraction and no signature derivation are involved. This is illustrated in Figure 3. With a proper choice of the image set mapping \( \phi \), what remains is to find an efficient way to compute \( d_\phi(A, B) \). As we shall see later, \( d_\phi(A, B) \) is indeed a pseudo distance over the set of all possible images when the image set mapping \( \phi \) and pixel level distortion function \( d(\cdot, \cdot) \) satisfy some symmetry condition.

The rest of the paper is organized as follows. Section II formally defines the SMID \( d_\phi(A, B) \), analyzes its distance property, and also presents some examples of \( \phi \). Section III establishes analytic formulas for and lower bounds to \( d_\phi(A, B) \) for some set mappings \( \phi \). In Section IV, the computation of \( d_\phi(A, B) \) according to its definition given in Section II is linked to the weighted bipartite matching problem for some \( \phi \), and a two-stage algorithm is presented to find an approximation for \( d_\phi(A, B) \) from above. In Section V, the SMID \( d_\phi(A, B) \) is computed with various block sizes, and are compared against popular signature based similarity functions, which demonstrates the superiority of the SMID \( d_\phi(A, B) \) over popular signature based similarity functions in terms of their consistency with human perceptual similarity. Finally, Section VI concludes the paper.

II. SMID DEFINITION AND ITS DISTANCE PROPERTIES

To simplify our discussion, we consider only gray-scale images. Let

\[
\mathcal{A} = \{A : A = (a(i, j)), -\infty < a(i, j) < +\infty, \\
1 \leq i \leq m, 1 \leq j \leq n \}
\]

denote the set of all gray-scale images of size \( n \times m \). Here each pixel is allowed to take any real value in the real line \( \mathbb{R} \).

Let \( \phi \) be a mapping which maps each image \( A \in \mathcal{A} \) to a non-empty set of images \( \phi(A) \subset \mathcal{A} \). Hereafter, such a mapping \( \phi \) will be referred to as a set mapping.

**Definition 1**: Give a set mapping \( \phi \) and pixel value distortion function \( d : \mathbb{R} \times \mathbb{R} \to [0, +\infty) \), the SMID between any two images \( A = (a(i, j)) \) and \( B = (b(i, j)) \), \( A, B \in \mathcal{A} \), with respect to \( \phi \) and \( d \) is defined as

\[
d_\phi(A, B) = \Delta \inf_{x \in \phi(A), y \in \phi(B)} d(X, Y)
\]

where

\[
d(X, Y) = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} d(x(i, j), y(i, j))
\]

for \( X = (x(i, j)) \) and \( Y = (y(i, j)) \).

Ideally, the image set \( \phi(A) \) for each image \( A \in \mathcal{A} \) would consist of all images in \( \mathcal{A} \) which are perceptually similar to \( A \). Of course, such an ideal image set \( \phi(A) \) is difficult, if not impossible, to construct. Instead, in what follows, we will give several exemplary mappings \( \phi \) which reflect some understandings on when humans would regard two images as similar.

**Example 1**: Let \( k \) and \( l \) be a factor of \( m \) and \( n \), respectively. For each image \( A = (a(i, j)) \in \mathcal{A} \), let \( \phi_0^{k,l}(A) \) be the set of all images \( X \in \mathcal{A} \) which are obtained from \( A \) by permuting a finite number of non-overlapping \( k \times l \)-blocks of \( A \). (Throughout the paper, a block of size \( k \times l \) (i.e., having \( k \) rows and \( l \) columns) is referred to as a \( k \times l \)-block.) Clearly, the motivation behind the definition of the set mapping \( \phi_0^{k,l}(A) \) is to reflect the understanding that if one moves objects in \( A \) around, the resulting image will often look similar to \( A \) to humans.

**Definition 2**: Given a set mapping \( \phi \), a binary relation \( \sim \) over \( \mathcal{A} \) is said to be induced by \( \phi \) if for any \( A, B \in \mathcal{A}, A \sim B \) if and only if \( B \in \phi(A) \). Denote such a binary relation by \( \sim_\phi \). A set mapping \( \phi \) is said to be transitive if its induced binary relation \( \sim_\phi \) is an equivalence relation.

**Theorem 1**: A set mapping \( \phi \) is transitive if and only if

(a) for any \( A \in \mathcal{A}, A \in \phi(A) \); and

(b) for any \( A, B \in \mathcal{A}, \phi(A) \) and \( \phi(B) \) are either the same or disjoint.

In particular, when \( \phi \) is transitive, we have \( \phi(B) = \phi(A) \) for any \( B \in \phi(A) \).

The proof of Theorem 1, along with other omitted proofs of results stated hereafter, can be found in the full paper [28].
is harder for non-transitive mappings \( \phi \) than transitive mappings \( \phi \) since characterizing a non-transitive mapping \( \phi \) itself is difficult in general. Below we present several transitive mappings \( \phi \).

**Example 2:** Let \( k \) and \( l \) be a factor of \( m \) and \( n \), respectively. For each image \( A = (a(i,j)) \in A \), partition it into non-overlapping \( k \times l \)-blocks; let \( \phi^{k,l}_i(A) \) be the set of all images \( X \in A \) which are obtained from \( A \) by permuting only those partitioned blocks in \( A \). When \( k \) and \( l \) are small in comparison with \( m \) and \( n \), \( \phi^{k,l}_i(A) \) still reflects the understanding that if one moves objects in \( A \) around, the resulting image will often look similar to \( A \) to humans.

**Example 3:** Let \( k \) and \( l \) be a factor of \( m \) and \( n \), respectively. An image \( P \in A \) is called a pattern image with pattern size \( k \times l \) if its \( k \times l \)-blocks with top left positions at \((ik + 1, jl + 1), i = 0, 1, 2, \ldots, j = 0, 1, 2, \ldots \) are all the same. For each image \( A = (a(i,j)) \in A \), let

\[
\phi^{k,l}_2(A) \triangleq \bigcup_{P} \phi^{k,l}_i(A + P)
\]

(2.3)

where the union is taken over all pattern images \( P \) with pattern size \( k \times l \), and the addition is the matrix addition with \( A \) and \( P \) regarded as \( m \times n \)-matrices. The inclusion of \( P \) in the definition of \( \phi^{k,l}_2(A) \) reflects the understanding that the superpositioned image \( A + P \) still looks similar to \( A \) to humans when \( k \) and \( l \) are small in comparison with \( m \) and \( n \).

**Theorem 2:** \( \phi^{k,l}_2 \) is transitive.

**Definition 3:** A set mapping \( \phi \) and pixel value distortion function \( d \) are said to satisfy a symmetry condition if for any \( A, B \in A \),

\[
\inf_{Y \in \phi(B)} d(X_1, Y) = \inf_{Y \in \phi(B)} d(X_2, Y)
\]

(2.4)

for any \( X_1, X_2 \in \phi(A) \).

**Theorem 3:** Let \( \phi \) and \( d \) be a set mapping and pixel value distortion function satisfying the symmetry condition (2.4). Suppose that there is a continuous, strictly increasing function \( h : [0, +\infty) \to [0, +\infty) \) such that \( h(d(X, Y)) \) is a distance over \( A \), where \( d(X, Y) \) is defined in (2.2). Then the following hold:

(a) \( h(d_{\phi}(A, B)) \) is a pseudo distance over \( A \).

(b) If \( \phi \) is also transitive and the infimum in (2.1) is achievable for any \( A, B \in A \), then

\[d_{\phi}(A, B) = 0 \]

if and only if \( B \in \phi(A) \).

**Remark 1:** The reader who is familiar with the Monge-Kantorovich (MK) distance [23] would recognize that the inner infimum in (3.3) is actually equal to the MK distance between a translated distribution \( \mu \) (with the origin translated to \( a^{k,l} \)) and a translated distribution \( \nu \) (with the origin translated to \( b^{k,l} \)) with respect to the pixel block distortion defined in (3.1) through the pixel value distortion function \( d \). However, with the outer infimum over all \( a^{k,l} \) and \( b^{k,l} \), the new quantity

**III. CHARACTERIZATION AND LOWER BOUNDS**

In this section, we will establish analytic formulas for and lower bounds to \( d_{\phi^{k,l}}(A, B) \) for any \( A, B \in A \). From now on, we identify a vector \( x^{k,l} = (x_1, x_2, \ldots, x_k) \in \mathbb{R}^{kl} \) with a \( k \times l \)-block or matrix and vice versa. Given a pixel value distortion function \( d : \mathbb{R} \times \mathbb{R} \to [0, +\infty) \), define for any \( x^{k,l} = (x_1, x_2, \ldots, x_k) \in \mathbb{R}^{kl} \) and \( y^{k,l} = (y_1, y_2, \ldots, y_l) \in \mathbb{R}^{kl} \)

\[
d(x^{k,l}, y^{k,l}) \triangleq \frac{1}{kl} \sum_{i=1}^{kl} |x_i - y_i|^r
\]

(3.1)

and

\[
|x^{k,l} - y^{k,l}|^r = \frac{1}{kl} \sum_{i=1}^{kl} |x_i - y_i|^r.
\]

For each image \( A \in A \), let \( P^{k,l}_A \) be the empirical distribution of all non-overlapping \( k \times l \)-blocks of \( A \) with top left positions at \((ik + 1, jl + 1), i = 0, 1, 2, \ldots, m/k - 1, j = 0, 1, 2, \ldots, n/l - 1 \). In addition, for any random variables \( X \) and \( Y \) taking values over \( \mathbb{R}^{kl} \), let \( P_X \) denote the marginal distribution of \( X \) over \( \mathbb{R}^{kl} \), and \( P_{XY} \) the joint distribution of \( X \) and \( Y \) over \( \mathbb{R}^{kl} \times \mathbb{R}^{kl} \), respectively. We then have the following analytic formulas for \( d_{\phi^{k,l}_2}(A, B) \).

**Theorem 4:** For any pixel value distortion function \( d : \mathbb{R} \times \mathbb{R} \to [0, +\infty) \) and any \( A, B \in A \),

\[
d_{\phi^{k,l}_2}(A, B) = \inf_{a^{k,l}, b^{k,l}} \min \{ E[d(X + a^{k,l}, Y + b^{k,l})] : P_{X} = P^{k,l}_A, P_{Y} = P^{k,l}_B, P_{XY} \text{ is an } \frac{m}{n} \text{-type} \} \tag{3.2}
\]

where the outer infimum in (3.2) is taken over all \( a^{k,l}, b^{k,l} \in \mathbb{R}^{kl} \), the inner minimization in (3.2) is taken over all jointly distributed random variables \( X, Y \) each taking values in \( \mathbb{R}^{kl} \) with their joint distribution satisfying the conditions in (3.2), and \( E[\cdot] \) stands for expectation.

Relaxing the type constraints in the inner minimization of (3.2) and defining for any probability distributions \( \mu \) and \( \nu \) over \( \mathbb{R}^{kl} \)

\[
T_d(\mu, \nu) \triangleq \inf_{a^{k,l}, b^{k,l}} \min \{ E[d(X + a^{k,l}, Y + b^{k,l})] : P_X = \mu, P_Y = \nu \} \tag{3.3}
\]

where the outer infimum is taken over all \( a^{k,l}, b^{k,l} \in \mathbb{R}^{kl} \), and the inner infimum is taken over all jointly distributed random variables \( X, Y \) each taking values in \( \mathbb{R}^{kl} \) with their marginal distributions \( P_X = \mu \) and \( P_Y = \nu \), we then get the following lower bounds.

**Corollary 2:** For any pixel value distortion function \( d : \mathbb{R} \times \mathbb{R} \to [0, +\infty) \) and any \( A, B \in A \),

\[
d_{\phi^{k,l}_1}(A, B) \geq T_d(F^{k,l}_A, F^{k,l}_B) \tag{3.4}
\]
$T_d(\mu, \nu)$ defined in (3.3) is not equal to the MK distance between $\mu$ and $\nu$ (with respect to the pixel block distortion) any more in general. For this reason, we shall coin a new name for the quantity $T_d(\mu, \nu)$ and refer to it as the translation invariant MK distance between $\mu$ and $\nu$ with respect to the pixel block distortion. In addition, although the concept of MK distance has been explored in image retrieval under the name of earth movers distance (EMD) [3], [16], it has been mainly applied to signatures of images, which are in turn summarizations of features such as colors, textures, shapes, and salient points extracted from images, with respect to some underlying distance between feature vectors [3], [16], [6], [17], [18]. With reference to Figure 2, as discussed in Section I, once images are abstracted into and represented by their signatures, the notion of distance between feature vectors and hence the resulting MK distance between signatures becomes absurd and less meaningful to the original images. On the other hand, the inner infimum of $T_d(P_{A_1}^{k,l}, P_{B_1}^{k,l})$ is related to the MK distance between the pixel block distributions of $A$ and $B$ with respect to the pixel block distortion defined in (3.1) through the pixel value distortion function $d$ and reflects the distance between two sets of images through the pixel value distortion function $d$. No feature extraction and no signature derivation are involved.

**Corollary 3:** When $d(x,y) = |x-y|^2$, we have for any images $A, B \in A$ and distributions $\mu$ and $\nu$

$$d_{d_{V^2}}(A,B) = \inf \{ |E[X] - E[Y]|^2 : P_X = P_{A_1}^{k,l}, P_Y = P_{B_1}^{k,l}, P_{XY} \text{ is an } \frac{mn}{kl} \text{-type} \} \quad (3.5)$$

and

$$T_d(\mu, \nu) = \inf \{ |E[X] - E[Y]|^2 : P_X = \mu, P_Y = \nu \}. \quad (3.6)$$

Since the gap $d_{d_{V^2}}(A,B) - T_d(P_{A_1}^{k,l}, P_{B_1}^{k,l})$ is in the order of $O(\frac{kl}{mn})$, one could use $T_d(P_{A_1}^{k,l}, P_{B_1}^{k,l})$ or its respective close lower bound as approximations for $d_{d_{V^2}}(A,B)$. There is an efficient way to compute a close lower bound of $T_d(P_{A_1}^{k,l}, P_{B_1}^{k,l})$, which will be used in our future work to determine if $A$ and $B$ could be regarded as similar. In the next section, however, we will compute $d_{d_{V^2}}(A,B)$ directly through its original definition (2.1) by establishing a link to weighted bipartite matching.

**IV. SMID CALCULATION AND WEIGHTED BIPARTITE MATCHING**

In this section, we consider only the squared error distortion $d(x,y) = |x-y|^2$. We first link the calculation of $d_{d_{V^2}}(A,B)$ to weighted bipartite matching [19], [20], and then present two algorithms to compute $d_{d_{V^2}}(A,B)$ from above. We begin with reviewing the weighted bipartite matching problem and its solutions in the literature.

**A. Review of weighted bipartite matching**

In graph theory, a weighted bipartite graph is a triple $G = (U \cup V, E, w)$, where $U$ and $V$ are two disjoint sets of vertices, $E$ is a subset of $U \times V$ representing all edges, and $w : E \to [0, +\infty)$ is a weight function assigning to each edge $e \in E$ a weight $w(e)$. A subset $F$ of $E$ is said to be a matching in $G$ if no two edges in $F$ are incident with a common vertex. A matching $F$ is perfect if every vertex in $G$ is incident to exactly one edge in $F$. The weight $w(F)$ of a matching $F$ is defined as the sum of weights of its edges, i.e., $w(F) = \sum_{e \in F} w(e)$. Given a weighted bipartite graph $G$, the minimum weight bipartite matching problem is to find a perfect matching in $G$ with minimum weight. Since only perfect matchings are compared, the existence of a perfect matching implies that $|U| = |V|$.

The minimum weight bipartite matching problem is also called the assignment problem in operations research, which has been well studied with solutions published starting from the 19th century by Carl Jacobi in Latin. A well-known solution, however, is the so-called Hungarian algorithm, which was invented in 1955 [14]. The original Hungarian algorithm has a time complexity $O(N^2 M)$, where $N$ is the total number of vertices in $G$ and $M$ is the number of edges in $G$. Later on, it was improved to achieve an $O(NM)$ running time by Edmonds [15]. A recent development on the Edmonds’ algorithm, named Blossom-V, was published in 2009 by incorporating both a variable $\delta$ approach and the use of priority queues [20]. Though the new development does not improve the worst-case complexity, it reduces the average running time in practice by an order of magnitude when compared with earlier implementations [20].

**B. Conversion into weighted bipartite matching**

Fix $k|m$ and $l|n$. Let us first demonstrate how to convert the calculation of $d_{d_{V^2}}(A,B)$ into the minimum weight bipartite matching problem for a weighted bipartite graph $G_{k,l}(A,B)$. Partition $A$ into non-overlapping $k \times l$-blocks, and write $A$ as $A = (A_{ij})_{1 \leq i \leq m/k, 1 \leq j \leq n/l}$, where $A_{ij}$ is a $k \times l$-block. The same notation applies to $B$ as well. Let $a_{kl}$ be the average block of $A_{ij}$, $1 \leq i \leq m/k$, $1 \leq j \leq n/l$, and $b_{kl}$ the average block of $B_{ij}$, $1 \leq i \leq m/k$, $1 \leq j \leq n/l$. That is, $a_{kl} = \frac{1}{k l} \sum_{i=1}^{m/k} \sum_{j=1}^{n/l} A_{ij}$ and $b_{kl} = \frac{1}{k l} \sum_{i=1}^{m/k} \sum_{j=1}^{n/l} B_{ij}$. Let $P^{(1)}$ and $P^{(2)}$ be two pattern images with pattern size $k \times l$ such that $P^{(1)} = (P^{(1)}_{ij})_{1 \leq i \leq m/k, 1 \leq j \leq n/l}$ and $P^{(2)} = (P^{(2)}_{ij})_{1 \leq i \leq m/k, 1 \leq j \leq n/l}$, where $P^{(1)}_{ij} = a_{kl}$ and $P^{(2)}_{ij} = b_{kl}$, $1 \leq i \leq m/k, 1 \leq j \leq n/l$. In other words, $P^{(1)}$ and $P^{(2)}$ are two pattern images with patterns $a_{kl}$ and $b_{kl}$, respectively. Define the weighted bipartite graph $G_{k,l}(A,B)$ associated with $A$ and $B$ as the complete bipartite graph $G_{k,l}(A,B) = (U \cup V, E, w)$ with $U = \{ u_{ij} : 1 \leq i \leq m/k, 1 \leq j \leq n/l \}$, $V = \{ v_{rs} : 1 \leq r \leq m/k, 1 \leq s \leq n/l \}$, $E = U \times V$, and $w$ assigning to each edge $e = (u_{ij}, v_{rs}) \in E$ a weight

$$w(e) = d(A_{ij} - P^{(1)}_{ij}, B_{rs} - P^{(2)}_{rs}).$$

In view of Corollary 3 and Definition 1, one can then verify that $\frac{mn}{kl}d_{d_{V^2}}(A,B)$ is equal to the minimum weight of perfect matchings in $G_{k,l}(A,B)$. Therefore, calculating $d_{d_{V^2}}(A,B)$ is equivalent to solving the minimum weight perfect matching problem for $G_{k,l}(A,B)$. In addition, it is easy to see that $G_{k,l}(A,B) = G_{k,l}(A - P^{(1)}, B - P^{(2)})$. 
C. Two-stage algorithm

By working with the weighted bipartite graph \(G_{k,l}(A, B)\), we can apply any suitable algorithm for weighted bipartite matching to compute or approximate \(d_{\phi^k_l}(A, B)\). The computation complexity would depend highly on the number \(N\) of vertices in \(U\), which in turn depends on the ratio between the size of \(k \times l\)-block and the image size in the case of \(G_{k,l}(A, B)\). For example, let us look at \(G_{k,l}(A, B)\) with image size \(256 \times 256\). When \(k = l = 8\), we have \(N = 1024\) and it takes a few seconds for the fast implementation of Blossom-V in C from [20] to compute \(d_{\phi^k_l}(A, B)\) and find the optimal perfect matching on a computer with Intel i7 CPU @ 2.8GHz. When \(k = l = 4\), we have \(N = 4096\) and the running time of the Blossom-V implementation increases to several minutes to half an hour on the same computer. When \(k = l = 2\), we have \(N = 16384\) and the running time of the Blossom-V implementation increases ever further. On the other hand, as shown in Section V, in order for \(d_{\phi^k_l}(A, B)\) to have good discriminating power for measuring image similarity, the block size \(k \times l\) can not be too large relatively to the image size. Therefore, directly applying existing weighted bipartite matching algorithms to the weighted complete bipartite graph \(G_{k,l}(A, B)\) with interesting block size \(k \times l\) may not be desirable for large scale image processing/management systems.

The above mentioned problem is partly due to the fact that \(G_{k,l}(A, B)\) is a complete bipartite graph and hence very dense (i.e., has too many edges). To alleviate this problem, we shall work instead with a weighted sparse, partial bipartite graph \(\tilde{G}_{k,l}(A, B)\) obtained by pruning edges from \(G_{k,l}(A, B)\). This leads to a generic two-stage algorithm which first constructs the weighted sparse bipartite graph \(\tilde{G}_{k,l}(A, B)\) at its stage one and then applies weighted bipartite matching algorithms such as Blossom-V to \(\tilde{G}_{k,l}(A, B)\) at its stage two. The resulting minimum weight of perfect matching in \(\tilde{G}_{k,l}(A, B)\) is equal to \(d_{\phi^k_l}(A, B)\) if all edges of the optimal perfect matching in \(G_{k,l}(A, B)\) appear in \(\tilde{G}_{k,l}(A, B)\), and an upper bound to \(d_{\phi^k_l}(A, B)\) otherwise.

There are several ways to construct the weighted sparse bipartite graph \(\tilde{G}_{k,l}(A, B)\). One way is to make use of a large block size. Select \(K\) and \(L\) such that \(K|n, L|n, k|K, l|L\). With \(P^{(1)}\) and \(P^{(2)}\) defined in Section IV.B, consider the weighted complete bipartite graph \(G_{K,L}(A - P^{(1)}, B - P^{(2)})\). Apply any minimum weight perfect matching algorithm such as Blossom-V to \(G_{K,L}(A - P^{(1)}, B - P^{(2)})\) to get an optimal perfect matching with minimum weight in \(G_{K,L}(A - P^{(1)}, B - P^{(2)})\), which is less computationally complex when \(K\) and \(L\) are large. Note that each vertex in \(G_{K,L}(A - P^{(1)}, B - P^{(2)})\) is corresponding to a \(K \times L\)-block in \(A\) or \(B\), and each vertex in \(G_{k,l}(A, B) = G_{K,L}(A - P^{(1)}, B - P^{(2)})\) is corresponding to a \(k \times l\)-block in \(A\) and \(B\). A vertex in \(G_{k,l}(A, B)\) is said to be contained in a vertex in \(G_{K,L}(A - P^{(1)}, B - P^{(2)})\) if the \(k \times l\)-block corresponding to the former is contained in the \(K \times L\)-block corresponding to the latter. An edge in \(G_{k,l}(A, B)\) is said to be feasible if its two vertices are contained in two vertices of \(G_{K,L}(A - P^{(1)}, B - P^{(2)})\) which are joined by an edge in the optimal perfect matching with minimum weight in \(G_{K,L}(A - P^{(1)}, B - P^{(2)})\). Then prune the bipartite graph \(G_{k,l}(A, B)\) by deleting all of its infeasible edges. One can then take the resulting pruned bipartite graph as \(\tilde{G}_{k,l}(A, B)\).

The two stage algorithm with such \(G_{k,l}(A, B)\) is summarized in Algorithm 1.

Algorithm 1 A two-stage algorithm for approximating \(d_{\phi^k_l}(A, B)\) from above.

1: Partition each of images \(A\) and \(B\) into non-overlapping \(k \times l\)-blocks. Compute their respective average blocks \(a^{k,l}\) and \(b^{k,l}\), and determine their respective pattern images \(P^{(1)}\) and \(P^{(2)}\) with patterns \(a^{k,l}\) and \(b^{k,l}\), respectively.
2: Select large \(K\) and \(L\) such that \(K | n, L | n, k | K, l | L\).
3: Build the weighted bipartite graph \(G_{K,L}(A - P^{(1)}, B - P^{(2)})\).
4: Apply any minimum weight perfect matching algorithm such as the Hungarian algorithm, Edmonds’ algorithm, or their variants to find an optimal perfect matching with minimum weight in \(G_{K,L}(A - P^{(1)}, B - P^{(2)})\).
5: Build the weighted sparse bipartite graph \(\tilde{G}_{k,l}(A, B)\) by pruning \(G_{k,l}(A, B)\) through deleting all of its infeasible edges.
6: Apply any minimum weight perfect matching algorithm to the sparse graph \(\tilde{G}_{k,l}(A, B)\) to find its optimal prefect matching with minimum weight. Then take the resulting normalized minimum weight as an approximation for \(d_{\phi^k_l}(A, B)\).

V. EXPERIMENTAL RESULTS

Experiments have been carried out to test the discriminating power of the proposed SMID for measuring image similarity and to illustrate the alignment between the proposed SMID and human perception for image similarity.

A. Experiment Settings

1) Testing Data Collection: As image similarity is mostly a subjective concept, it is critical to establish a testing set with human inputs. For that purpose, we set up an online voting system [21]. Selected images come from two sources, including the INRIA Holidays dataset [22] and a set of our own photo collections over the past few years. Since similarity is measured between a pair of images and there are a lot more dissimilar pairs than similar pairs, one strategy we used to ensure a reasonably-balanced data set is to apply a time-adjacent criterion, i.e., a pair of images will be included only when their creation time is within two-hours apart. This criterion helps to balance the number of similar pairs and the number of dissimilar pairs.

To further avoid uncertainty in the collected human inputs on image similarity, we only select 1116 pairs to use, for which all voters reach a consensus on whether they are similar or dissimilar. In other words, all pairs with ambiguity—either they are categorized as uncertain by at least one person or there is any discrepancy among all voters—are excluded from the following experiment. As a result, there are 550 pairs of similar photos and 566 pairs of dissimilar photos. In addition,
a resizing procedure is applied to each photo to normalize the image resolution. Specifically, for an image with resolution $n \times m$, it is resized to the resolution of $256\sqrt{n/m} \times 256\sqrt{m/n}$.

2) Comparative study setting: We first want to assess the impact of block size $k \times l$ on the discriminating power of our proposed SMID for measuring image similarity by comparing the SMID metrics based on different block sizes $k \times l = 1 \times 1, 2 \times 2, 4 \times 4, 8 \times 8$. We then select one block size $k \times l$, which offers a good trade-off between discriminating power and computational complexity relative to the image size $256\sqrt{n/m} \times 256\sqrt{m/n}$, and compare the resulting SMID with other image similarity metrics based on image signatures.

The comparison among the SMID metrics based on different block sizes $k \times l$ relates to three types of methods for computing the SMID: (1) a direct computation of weighted bipartite matching, (2) an approximation by Algorithm 1, and (3) a closed-form calculation in the special case with $k = l = 1$ as given by (3.6). Note that the graph $\hat{G}_{k,l}(A, B)$ in Algorithm 1 consists of $\frac{m}{k^2}$ disconnected sub-graphs, each with $\frac{2k^2}{k^2} \cdot L$ nodes. Thus, the matching in Step 6 of Algorithm 1 can be dramatically sped up by solving each sub-graph independently.

Comparison between the SMID based on the selected block size $k \times l$ and other similarity metrics based on image signatures is carried out by studying four popular image similarity metrics: histogram intersection (HI) [5], [4], the MK distance [23], [16], scale invariant feature transform (SIFT) [6], [7], [8], and autocorrelogram [11].

The HI distance between two histograms $A$ and $B$ as $A = a_1, a_2, \cdots, a_m$ and $B = b_1, b_2, \cdots, b_m$ is defined as

$$K(A, B) = \sum_{i=1}^{m} \min(a_i, b_i)$$

where $m$ is the number of bins in the histograms. Essentially, it reflects the global color statistics of images, while overlooking the shape and texture of images.

Note that the HI distance depends only on two color distributions and has nothing to do with the underlying distance between colors. On the other hand, the MK distance, also referred to as EMD in computer science, involves both distributions and the underlying distance between colors or feature vectors by considering minimum cost mass transportation between two distributions with the underlying distance as a transportation cost. Consequently, it shows certain robustness against image blurring, dithering or other local deformations and is widely used in content-based image retrieval [17].

The SIFT represents another type of similarity metrics that focus on local features/Textures detection and characterization [6], [7]. It has been regarded as one of the most pervasive and robust image features [9] for image similarity assessment. For the purpose of comparison, here we adopt the SIFT implementation from [8]. The original SIFT assessment was further improved in [10] by introducing the EMD for matching descriptors with the underlying $L_1$ ground distance. For completeness, we also include the EMD-based SIFT similarity assessment in our comparison.

Autocorrelogram is another popular technique for similarity measurement in content-based image retrieval applications [11]. In our implementation for comparison, the autocorrelogram metric is computed according to Eq. (3) in [11] by using a distance set $D = \{1, 3, 5, 7\}$ with each pixel quantized into 64 values.

3) Testing criterion: Since our testing data set is labeled without ambiguity, an effective way to demonstrate the power of a similarity metric to discriminate between pairs of similar images and pairs of dissimilar images is to plot the distributions of the similarity metric over the set of all pairs of similar images and the set of all pairs of dissimilar images, respectively. If the similarity metric is single-valued—the SMID, MK distance, HI metric, and autocorrelogram metric are all single-valued—these two distributions are one dimensional. In this case, if the two distributions do not overlap, it implies that the similarity metric can discriminate between pairs of similar images and pairs of dissimilar images perfectly without any error. Otherwise, the receiver operating characteristic (ROC) curve will be used to further illustrate the overlapping region. In determining the ROC curve, we consider a simple discriminator which would classify a pair of images as similar if and only if its similarity metric value is less than or equal to $\alpha$ in the case of the SMID, MK distance, or autocorrelogram metric, and greater than or equal to $\alpha$ in the case of the HI metric, where $\alpha$ is in the overlapping region. A false positive event happens whenever this discriminator classifies a pair of dissimilar images as similar.

If the similarity metric is not single-valued, then the distributions of the similarity metric over the set of all pairs of similar images and the set of all pairs of dissimilar images are multiple dimensional. Both the SIFT and EMD-based-SIFT are represented by two values: the number of matches (denoted by $nm$) and the average distance of all matches (denoted by $emd$ in the case of the SIFT and by $emd$ in the case of the EMD-based SIFT). Therefore, their corresponding distributions are two dimensional.

A single-valued similarity metric $\beta$ is said to be consistent with human perception for image similarity if for any images $A$, $B$, and $C$, one has $\beta(A, B) < \beta(A, C)$ whenever images $A$ and $B$ are deemed by human to be more similar to each other than images $A$ and $C$. A good similarity metric not only possesses good power to discriminate between pairs of similar images and pairs of dissimilar images, but also should be consistent with human perception for image similarity. Therefore, when we compare the SMID with the MK distance, HI metric, autocorrelogram metric, SIFT, and EMD-based SIFT, it is also interesting to test their respective consistency with human perception for image similarity. For this purpose, a small set of photos was selected from [22], which contains five (5) photos with names ‘124200.jpg’, ‘124201.jpg’, ‘124202.jpg’, ‘124210.jpg’, and ‘123601.jpg’. For convenience, hereafter, these five photos are referred to as Img1, Img2, Img3, Img4, and Img5, respectively. The Img1 will be used as a reference, and the other four photos will be sorted according to their similarity distances from Img1 calculated according to a given similarity metric when the similarity metric is tested for its consistency with human perception.
B. Experiment Results

1) SMID with various block sizes: Since each original image of large size $n \times m$ was down sampled to the size $256\sqrt{n/m} \times 256\sqrt{m/n}$, four different block sizes were selected for testing: $k \times l = 1 \times 1, 2 \times 2, 4 \times 4, 8 \times 8$. Figure 4 shows the distributions of their respective SMID metrics over the set of all pairs of similar images and the set of all pairs of dissimilar images, where $T_d(P_{A,1}^1, P_{B,1}^1)$ is computed using (3.6), $d_{8x8}(A, B)$ is computed using Algorithm 1, and $d_{2x2}(A, B)$ and $d_{4x4}(A, B)$ are computed directly as a weighted bipartite matching problem using a Hungarian/Edmond algorithm. From Figure 4, it follows that the discriminating power of the SMID degrades when the block size is either too small or too large relative to the down-sampled image size $256\sqrt{n/m} \times 256\sqrt{m/n}$—the overlapping area between the two distributions is larger in the case of $k \times l = 1 \times 1$ or $8 \times 8$ than in the case of $k \times l = 2 \times 2$ or $4 \times 4$. The ROC curves in Figure 5 further reinforce this, clearly showing that the SMID with block size $2 \times 2$ or $4 \times 4$ significantly outperforms the SMID with block size $1 \times 1$ or $8 \times 8$ in terms of discriminating between pairs of similar images and pairs of dissimilar images. Furthermore, note that the $2 \times 2$ results $d_{2x2}(A, B)$ are obtained by Algorithm 1, which is much faster\(^1\) than the computation of $d_{8x8}(A, B)$. Therefore, for image size $256\sqrt{n/m} \times 256\sqrt{m/n}$, the block size $k \times l = 2 \times 2$ with $d_{2x2}(A, B)$ computed by Algorithm 1 seems to offer a good trade-off between similarity discriminating power and computational complexity. In the following, $d_{2x2}(A, B)$ computed by Algorithm 1 will be used for comparing with other metrics based on signatures.

2) Discriminating power comparison between SMID and other metrics: Figure 6 compares the SMID $d_{2x2}(A, B)$ (computed by Algorithm 1) with three other single-valued similarity metrics: the HI metric, autocorrelogram metric, and MK distance. The corresponding ROC curves for these four metrics are shown in Figure 7. It is clear to see that the SMID $d_{2x2}(A, B)$ outperforms significantly the HI metric, autocorrelogram metric, and MK distance in terms of discriminating between pairs of similar images and pairs of dissimilar images.

\(^1\)The computation of $d_{2x2}(A, B)$ by an Edmond/Hungarian algorithm is about 100 times slower than computation of $d_{2x2}(A, B)$ by Algorithm 1.

Fig. 4. Distributions of SMID metrics with various block sizes over 550 pairs of similar images and 566 pairs of dissimilar images.

Fig. 5. ROC curves corresponding to Figure 4.

Fig. 6. Distributions of the SMID $d_{2x2}(A, B)$ (computed by Algorithm 1) and other single valued similarity metrics over 550 pairs of similar images and 566 pairs of dissimilar images.

C. Consistency comparison

Figure 10 shows the result of sorting four photos, Img2, Img3, Img4, and Img5, according to their SMID distances $d_{2x2}(A, B)$ from the reference Img1. Examine the five photos all together. It is not hard to obtain a consensus that Img2 and Img3 are more similar to the reference Img1 than Img4 and
Img4 are mistakenly believed by EMD-based SIFT to be more similar to the reference than Img3 is, because the former has more and stronger matches with the reference than the latter does.

Table I lists the results of sorting the four photos according to their similarity distances from the reference Img1 based on other five similarity metrics: the HI metric, autocorrelogram metric, MK distance, SIFT metric, and EMD-based SIFT metric, respectively. In Table I, underneath the name of each photo is its similarity value from the reference photo Img1 calculated according to the corresponding similarity metric, which is single valued in the case of the HI metric, autocorrelogram metric, and MK distance, and 2D vector valued in the case of SIFT and EMD-based SIFT. Photos are sorted in the increasing order of the similarity distance in the case of the autocorrelogram metric and MK distance, in the decreasing order of the similarity value in the case of HI, and in the decreasing order of \( nm/dm \) and \( nm/\text{emd} \) in the case of SIFT and EMD-based SIFT, respectively. For each row of Table I except for the first row, a photo with a lower rank value is believed, according to the corresponding similarity metric, to be more similar to the reference than a photo with a higher rank value is. With reference to Figure 10, it is clear that none of these five similarity metrics is consistent with human perception for image similarity. Specifically, Img4 or Img5 is mistakenly believed by the HI to be more similar to the reference than Img2 is; both the autocorrelogram metric and MK distance mistakenly choose Img4 over Img2 as the one more similar to the reference; SIFT mistakenly chooses Img4 as the one most similar to the reference because it possesses the largest number of matches with the strongest average strength of the matches among the four photos; and finally, Img5 are, though it is arguable that between Img2 and Img3 or between Img4 and Img5 which one is closer to the reference.

Figure 10 clearly shows that Img2 and Img3 are more similar to Img1 than Img4 and Img5 are. More specifically, Figure 10 shows Img3 is slightly more similar to the reference than Img2 is, which may be justified by zooming into the background behind the boats. Apparently, the background behind the boat of Img3 is closer to that in the reference than the background of Img2 is. Thus, the SMID \( d_{\phi^2}(A,B) \) is consistent with human perception for image similarity, at least for the five testing photos.

The derivation of the lower bound, its computation, and its lower bound to the SMID can also be efficiently calculated. Such a power the lower bound possesses for image similarity. Such a metric and human perception.

**D. Discussion**

From the above experimental results, it is fair to say that in comparison with other similarity metrics based on signatures, the SMID based on a properly chosen block size \( k \times l \) has better power to discriminate between pairs of similar images and pairs of dissimilar images and is better aligned with human perception for image similarity. Nonetheless, applying the SMID to large scale image management system may result in increased computational complexity. An alternate process may be to approach the SMID from below, as mentioned in Section III. Figure 12 shows the results obtained for the distributions of a lower bound of equation (3.6) with \( k = l = 2 \). As shown in Figure 12, the two distributions of such a lower bound do not even overlap, indicating the excellent discrimination power the lower bound possesses for image similarity. Such a lower bound to the SMID can also be efficiently calculated. The derivation of the lower bound, its computation, and its...
In this paper, a new image similarity metric dubbed the set mapping induced similarity distance (SMID) has been proposed. To determine the SMID between two images \( A \) and \( B \), one first maps each image \( A \) into a set \( \phi(A) \) of images, and then calculates the SMID between \( A \) and \( B \) as the smallest average distortion per pixel between any pair of images, one from \( \phi(A) \) and the other from \( \phi(B) \). Several examples of the set mapping \( \phi \) have been illustrated. The distance property of the SMID has been analyzed. Analytic formulas for and lower bounds to the SMID have also been established. The computation of the SMID according to its definition for some set mappings \( \phi \) has been linked to weighted bipartite matching and a two-stage algorithm has been developed for approximating the SMID efficiently. Experimental results have demonstrated that in comparison with existing popular similarity metrics in the literature such as histogram intersection, SIFT, the MK distance, and autocorrelogram, the SMID has better power to discriminate between pairs of similar images and pairs of dissimilar images and is better aligned with human perception for image similarity.

VI. CONCLUSION

In this paper, a new image similarity metric dubbed the set mapping induced similarity distance (SMID) has been proposed. To determine the SMID between two images \( A \) and \( B \), one first maps each image \( A \) into a set \( \phi(A) \) of images, and then calculates the SMID between \( A \) and \( B \) as the smallest average distortion per pixel between any pair of images, one from \( \phi(A) \) and the other from \( \phi(B) \). Several examples of the set mapping \( \phi \) have been illustrated. The distance property of the SMID has been analyzed. Analytic formulas for and lower bounds to the SMID have also been established. The computation of the SMID according to its definition for some set mappings \( \phi \) has been linked to weighted bipartite matching and a two-stage algorithm has been developed for approximating the SMID efficiently. Experimental results have demonstrated that in comparison with existing popular similarity metrics in the literature such as histogram intersection, SIFT, the MK distance, and autocorrelogram, the SMID has better power to discriminate between pairs of similar images and pairs of dissimilar images and is better aligned with human perception for image similarity.


